

# Convergence Analysis of Policy Iteration

Ali Heydari<sup>1</sup>

## Abstract

Adaptive optimal control of nonlinear dynamic systems with deterministic and known dynamics under a known undiscounted infinite-horizon cost function is investigated. Policy iteration scheme initiated using a stabilizing initial control is analyzed in solving the problem. The convergence of the iterations and the optimality of the limit functions, which follows from the established uniqueness of the solution to the Bellman equation, are the main results of this study. Furthermore, a theoretical comparison between the speed of convergence of policy iteration versus value iteration is presented. Finally, the convergence results are extended to the case of multi-step look-ahead policy iteration.

## I. INTRODUCTION

This short study investigates the convergence of the policy iteration (PI) as one of the schemes in implementation of adaptive/approximate dynamic programming (ADP), sometimes referred to by reinforcement learning (RL) or neuro-dynamic programming (NDP), [1]- [11].

Compared to its alternative, i.e., value iteration (VI), the PI calls for a higher computational load per iteration, due to a ‘full backup’ as opposed to a ‘partial backup’ in VI, [12]. However, the PI has the advantage that the control under evolution remains stabilizing [10], hence, it is more suitable for online implementation, i.e., adapting the control ‘on the fly’. The convergence analyses for PI with continuous state and control spaces and an undiscounted cost function are given in [10]. The results presented in this study however, are from a different viewpoint with different assumptions and lines of proofs. Moreover, interested readers are referred to the results from a simultaneous research (at least in terms of the availability of the results to the public) presented in [13], which are the closest to the first two theorems of this study.

This study establishes the convergence of the PI to the solution to the optimal control problem with known deterministic dynamics. Moreover, given the faster convergence of PI compared with VI which can be observed in numerical implementations, some theoretical results are presented which compare the rates of convergences. Finally, the multi-step look-ahead variation of PI, [14], is analyzed and its convergence is established.

## II. PROBLEM FORMULATION

The discrete-time nonlinear system given by

$$x_{k+1} = f(x_k, u_k), k \in \mathbb{N}, \quad (1)$$

is subject to control, where (possibly discontinuous) function  $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  is known, the state and control vectors are denoted with  $x$  and  $u$ , respectively, and  $f(0, 0) = 0$ . Positive integers  $n$  and  $m$  denote the dimensions of the continuous state space  $\mathbb{R}^n$  and the (possibly discontinuous) control space  $\mathcal{U} \subset \mathbb{R}^m$ , respectively, sub-index  $k$  represents the discrete time index, and the set of non-negative integer numbers is denoted with  $\mathbb{N}$ . The cost function subject to minimization is given by

$$J = \sum_{k=0}^{\infty} U(x_k, u_k), \quad (2)$$

where the utility function  $U : \mathbb{R}^n \times \mathcal{U} \rightarrow \mathbb{R}_+$  is positive semi-definite with respect to the first input, and positive definite with respect to its second input. The set of non-negative real numbers is denoted with  $\mathbb{R}_+$ .

Selecting an initial feedback control policy  $h : \mathbb{R}^n \rightarrow \mathcal{U}$ , i.e.,  $u_k = h(x_k)$ , the adaptive optimal control problem is updating/adapting the control policy such that cost function (2) is minimized. The minimizing control policy is called the optimal control policy and denoted with  $h^*(\cdot)$ .

**Notation 1.** The state trajectory initiated from the initial state  $x_0$  and propagated using the control policy  $h(\cdot)$  is denoted with  $x_k^h, \forall k \in \mathbb{N}$ . In other words,  $x_0^h := x_0$  and  $x_{k+1}^h = f(x_k^h, h(x_k^h)), \forall k \in \mathbb{N}$ .

**Definition 1.** The control policy  $h(\cdot)$  is defined to be asymptotically stabilizing within a domain if  $\lim_{k \rightarrow \infty} x_k^h = 0$ , for every initial  $x_0^h$  selected within the domain, [15].

**Definition 2.** The set of admissible control policies (within a compact set), denoted with  $\mathcal{H}$ , is defined as the set of policies  $h(\cdot)$  that asymptotically stabilize the system within the set and their respective ‘cost-to-go’ or ‘value function’, denoted with  $V_h : \mathbb{R}^n \rightarrow \mathbb{R}_+$  and defined by

$$V_h(x_0) := \sum_{k=0}^{\infty} U(x_k^h, h(x_k^h)), \quad (3)$$

<sup>1</sup> Assistant Professor of Mechanical Engineering, South Dakota School of Mines and Technology, Rapid City, SD 57701, email: ali.heydari@sdsmt.edu.

is upper bounded within the compact set by a continuous function  $\bar{V} : \mathbb{R}^n \rightarrow \mathbb{R}_+$  where  $\bar{V}(0) = 0$ .

If the value function itself is continuous, the upper boundedness by  $\bar{V}(\cdot)$  is trivially met, through selecting  $\bar{V}(\cdot) = V_h(\cdot)$ . Note that the continuity of the upper bound in the compact set leads to its finiteness within the set, and hence, the finiteness of the value function. This is a critical feature for the value function and hence, the control policy.

**Assumption 1.** *There exists at least one admissible control policy for the given system within a connected and compact set  $\Omega \subset \mathbb{R}^n$  containing the origin.*

**Assumption 2.** *The intersection of the set of  $n$ -vectors  $x$  at which  $U(x, 0) = 0$  with the invariant set of  $f(\cdot, 0)$  only contains the origin.*

Assumption 1 leads to the conclusion that the value function associated with the *optimal* control policy is finite at any point in  $\Omega$ , as it will not be greater than  $V_h(\cdot)$  at that point, for any admissible control policy  $h(\cdot)$ . Assumption 2 implies that the optimal control policy will be asymptotically stabilizing, as there is no non-zero state trajectory that can ‘hide’ somewhere without convergence to the origin. Given these two assumptions, it is concluded that the optimal control policy is an admissible policy, i.e.,  $h^*(\cdot) \in \mathcal{H}$ .

### III. ADP-BASED SOLUTIONS

The Bellman equation [4], given below, provides the *optimal value function*

$$V^*(x) = \min_{u \in \mathcal{U}} \left( U(x, u) + V^*(f(x, u)) \right), \quad (4)$$

which once obtained, leads to the solution to the problem, through

$$h^*(x) = \arg \min_{u \in \mathcal{U}} \left( U(x, u) + V^*(f(x, u)) \right). \quad (5)$$

But, this is mathematically impracticable for general nonlinear systems, [4]. Policy iteration (PI) provides a learning algorithms for training a function approximator or forming a lookup table, for approximating the solution [16], [14], [17]. This approximation is done within a compact and connected set, containing the origin, called the domain of interest and denoted with  $\Omega$ .

Starting with an initial admissible control policy, denoted with  $h^0(\cdot)$ , one iterates through the *policy evaluation equation* given by

$$V^i(x) = U(x, h^i(x)) + V^i(f(x, h^i(x))), \forall x \in \Omega, \quad (6)$$

and the *policy update equation* given by

$$h^{i+1}(x) = \arg \min_{u \in \mathcal{U}} \left( U(x, u) + V^i(f(x, u)) \right), \forall x \in \Omega, \quad (7)$$

for  $i = 0, 1, \dots$  until they converge, in PI. Each of these equations may be evaluated at different points in  $\Omega$ , for obtaining the *targets* for training the respective function approximators.

### IV. CONVERGENCE ANALYSIS OF POLICY ITERATION

Given the fact that Eqs. (6) and (7) are iterative equations, the following questions arise. 1- Does the iterations converge? 2- If they converge, are the limit functions optimal? This section is aimed at answering these two questions. Initially the following two lemmas are presented.

**Lemma 1.** *Given admissible control policies  $h(\cdot)$  and  $g(\cdot)$ , if*

$$U(x, h(x)) + V_g(f(x, h(x))) \leq V_g(x), \forall x \in \Omega, \quad (8)$$

*then  $V_h(x) \leq V_g(x), \forall x \in \Omega$ .*

*Proof:* Evaluating (8) at  $x_0^h \in \Omega$ , one has

$$U(x_0^h, h(x_0^h)) + V_g(x_1^h) \leq V_g(x_0^h), \forall x_0^h \in \Omega. \quad (9)$$

Also, evaluating (8) at  $x_1^h$  leads to

$$U(x_1^h, h(x_1^h)) + V_g(x_2^h) \leq V_g(x_1^h), \forall x_1^h \in \Omega. \quad (10)$$

Using (10) in (9) leads to

$$U(x_0^h, h(x_0^h)) + U(x_1^h, h(x_1^h)) + V_g(x_2^h) \leq V_g(x_0^h), \forall x_0^h \in \Omega. \quad (11)$$

Repeating this process for  $N - 2$  more times leads to

$$\sum_{k=0}^{N-1} U(x_k^h, h(x_k^h)) + V_g(x_N^h) \leq V_g(x_0^h), \forall x_0^h \in \Omega. \quad (12)$$

Letting  $N \rightarrow \infty$  and given  $V_g(x) \geq 0, \forall x$ , which hence can be dropped from the left hand side, inequality (12) leads to  $V_h(x) \leq V_g(x), \forall x \in \Omega$ , by definition of  $V_h(\cdot)$ .  $\square$

Assuming two admissible control policies  $h(\cdot)$  and  $g(\cdot)$ , Lemma 1 simply shows that if applying  $h(\cdot)$  at the first time step and applying  $g(\cdot)$  for infinite number of times in the future, leads to a cost not greater than only applying  $g(\cdot)$ , then, the value function of  $h(\cdot)$  also will not be greater than that of  $g(\cdot)$ , at any point.

**Lemma 2.** *Given admissible control policies  $h(\cdot)$  and  $g(\cdot)$ , if*

$$V_h(x) < V_g(x), \exists x \in \Omega, \quad (13)$$

*then*

$$U(x, h(x)) + V_g(f(x, h(x))) < V_g(x), \exists x \in \Omega. \quad (14)$$

*Proof:* The proof is done by contradiction. Assume that (14) does not hold, i.e.,

$$U(x_0^h, h(x_0^h)) + V_g(x_1^h) \geq V_g(x_0^h), \forall x_0^h \in \Omega, \quad (15)$$

which leads to

$$U(x_1^h, h(x_1^h)) + V_g(x_2^h) \geq V_g(x_1^h), \forall x_1^h \in \Omega. \quad (16)$$

Using (16) in (15) leads to

$$U(x_0^h, h(x_0^h)) + U(x_1^h, h(x_1^h)) + V_g(x_2^h) \geq V_g(x_0^h), \forall x_0^h \in \Omega. \quad (17)$$

Repeating this process for  $N - 2$  more times, one has

$$\sum_{k=0}^{N-1} U(x_k^h, h(x_k^h)) + V_g(x_N^h) \geq V_g(x_0^h), \forall x_0^h \in \Omega. \quad (18)$$

Let  $N \rightarrow \infty$ . Given the admissibility of  $g(\cdot)$  and  $h(\cdot)$ , one has  $V_g(x_N^h) \rightarrow 0, \forall x_0^h$ , as  $N \rightarrow \infty$ . The reason is  $\lim_{N \rightarrow \infty} x_N^h \rightarrow 0$  and the continuity of the upper bound of  $V_g(\cdot)$ , per the admissibility of  $g(\cdot)$ . Therefore, Inequality (18) contradicts (13), because, the second term in the left hand side of (18) can be made arbitrarily small. Hence, (18) leads to  $V_h(x) \geq V_g(x), \forall x \in \Omega$ <sup>1</sup>, which contradicts (13), hence, (15) cannot hold. This completes the proof.  $\square$

In simple words, Lemma 2 shows that if the value function of  $h(\cdot)$  is less than that of  $g(\cdot)$  at least at one  $x$ , then, the cost of applying  $h(\cdot)$  only at the first step and applying  $g(\cdot)$  for the rest of the steps also will be less than the cost of only applying  $g(\cdot)$  throughout the horizon, at least at one  $x$ . This result leads to the uniqueness of the solution to the Bellman equation (4), as shown in the next theorem.

**Theorem 1.** *The Bellman equation given by (4) has a unique solution in  $\Omega$ .*

*Proof:* The proof is by contradiction. Assume that there exists some  $V_h(\cdot)$  that satisfies

$$V_h(x) = \min_{u \in \mathcal{U}} (U(x, u) + V_h(f(x, u))), \forall x \in \Omega, \quad (19)$$

while,  $V^*(x) < V_h(x), \exists x \in \Omega$ , in other words  $h^*(x) \neq h(x), \exists x \in \Omega$ , where

$$h(x) := \arg \min_{u \in \mathcal{U}} (U(x, u) + V_h(f(x, u))), \forall x \in \Omega. \quad (20)$$

Using Lemma 2, inequality  $V^*(x) < V_h(x), \exists x \in \Omega$ , leads to

$$U(x, h^*(x)) + V_h(f(x, h^*(x))) < V_h(x) = U(x, h(x)) + V_h(f(x, h(x))), \exists x \in \Omega. \quad (21)$$

But, (21) contradicts (20). Hence,  $h^*(x) = h(x), \forall x \in \Omega$ , and therefore,  $V^*(x) = V_h(x), \forall x \in \Omega$ .  $\square$

The next step is the proof of convergence of PI.

**Theorem 2.** *The policy iteration given by equations (6) and (7) converges monotonically to the optimal solution in  $\Omega$ .*

*Proof:* The first step is showing the monotonicity of the sequence of value functions  $\{V^i(x)\}_{i=0}^{\infty}$  generated using the PI equations. By (7), one has

$$U(x, h^{i+1}(x)) + V^i(f(x, h^{i+1}(x))) \leq V^i(x), \forall x \in \Omega. \quad (22)$$

Using Lemma 2, the former inequality leads to

$$V^{i+1}(x) \leq V^i(x), \forall x \in \Omega, \quad (23)$$

<sup>1</sup>This conclusion can also be made using another contradiction argument, through (13) which leads to  $V_h(x_0) + \epsilon = V_g(x_0), \exists x_0 \in \Omega$ , for some  $\epsilon = \epsilon(x_0) > 0$ . Then, selecting large enough  $N$  such that  $V_g(x_N^h) < \epsilon$ , inequality (18) contradicts (13).

for any selected  $i$ . Hence,  $\{V^i(x)\}_{i=0}^\infty$  is pointwise decreasing. On the other hand, it is lower bounded by the optimal value function. Hence, it converges, [18]. Denoting the limit value function and the limit control policy with  $V^\infty(\cdot)$  and  $h^\infty(\cdot)$ , respectively, they satisfy PI equations

$$V^\infty(x) = U(x, h^\infty(x)) + V^\infty(f(x, h^\infty(x))), \forall x \in \Omega, \quad (24)$$

and

$$h^\infty(x) = \arg \min_{u \in \mathcal{U}} (U(x, u) + V^\infty(f(x, u))), \forall x \in \Omega, \quad (25)$$

hence,

$$V^\infty(x) = \min_{u \in \mathcal{U}} (U(x, u) + V^\infty(f(x, u))), \forall x \in \Omega. \quad (26)$$

Eq. (26) is the Bellman equation, which per Theorem 1 has a unique solution. Hence,  $V^\infty(\cdot) = V^*(\cdot)$  everywhere in  $\Omega$ . This completes the proof.  $\square$

**Theorem 3.** *The control policies at the iterations of the policy iteration given by equations (6) and (7) remain admissible in  $\Omega$ .*

*Proof:* Given the requirements for admissibility, one needs to show that each policy is asymptotically stabilizing and its respective value function is upper bounded by a continuous function which passes through the origin. The latter follows from the monotonicity of the sequence of value functions under VI, established in Theorem 2, since  $h^0(\cdot)$  is admissible. The former, also follows from this monotonicity, as no state trajectory can hide in the set at which the utility function is zero, without convergence to the origin, per Assumption 2. In other words, in order for its value function to be bounded, the policy needs to steer the trajectory towards the origin.  $\square$

## V. COMPARISON BETWEEN POLICY AND VALUE ITERATIONS

Value iteration (VI), as an alternative to PI, is conducted using an initial guess  $W^0(\cdot)$  and iterating through the *policy update equation* given by

$$g^i(x) = \arg \min_{u \in \mathcal{U}} (U(x, u) + W^i(f(x, u))), \forall x \in \Omega, \quad (27)$$

and the *value update equation*

$$W^{i+1}(x) = U(x, g^i(x)) + W^i(f(x, g^i(x))), \forall x \in \Omega. \quad (28)$$

The two former equations can be merged into

$$W^{i+1}(x) = \min_{u \in \mathcal{U}} (U(x, u) + W^i(f(x, u))), \forall x \in \Omega. \quad (29)$$

for  $i = 0, 1, \dots$ , where notations  $W^i(\cdot)$  and  $g^i(\cdot)$  are used for the value function and the control policy resulting from the VI, respectively, for clarity. The convergence proof of VI is not the subject of this study and can be found in many references including [19], [8], and [11].

The VI has the advantage of not requiring an admissible control as the initial guess. The PI, however, has the advantage that the control policies subject to evolution remain stabilizing for the system. It was shown in [20] that if the VI is also initiated using an admissible initial guess, the control policies remain stabilizing. Therefore, starting with an admissible guess, the VI and PI seem to be similar in terms of stability. The computational load per iteration in VI is significantly less than that of PI, due to needing to do a simple recursion in VI using (28), called a ‘partial backup’ in [12], as compared with solving an equation in PI, namely, Eq. (6), which is a ‘full backup’, [12]. However, in practice, it can be seen that the PI converges much faster than the VI, in terms of the number of iterations. This section is aimed at providing some analytical results confirming this observation.

**Theorem 4.** *If  $V^0(\cdot) = W^0(\cdot)$  is calculated using as admissible control policy, the policy iteration given by equations (6) and (7) converges not slower than the value iteration given by equations (27) and (28), in  $\Omega$ .*

*Proof:* Given the convergence of both schemes to the unique  $V^*(\cdot)$ , the claim is proved by showing that  $V^i(x) \leq W^i(x)$ ,  $\forall x \in \Omega$ ,  $\forall i \in \mathbb{N}$ . From  $V^0(\cdot) = W^0(\cdot)$  one has  $h^1(\cdot) = g^0(\cdot)$ . Hence,

$$\begin{aligned} W^1(x) &= U(x, g^0(x)) + W^0(f(x, g^0(x))) = U(x, h^1(x)) + V^0(f(x, h^1(x))) \geq \\ &U(x, h^1(x)) + V^1(f(x, h^1(x))) = V^1(x), \forall x \in \Omega, \end{aligned} \quad (30)$$

where the inequality is due to the monotonically decreasing nature of  $\{V^i(x)\}_{i=0}^\infty$  established in Theorem 2. Therefore,

$W^1(x) \geq V^1(x), \forall x$ . Now, assume that  $W^i(x) \geq V^i(x), \forall x$ , for some  $i$ . Then,

$$\begin{aligned} W^{i+1}(x) &= U(x, g^i(x)) + W^i(f(x, g^i(x))) \geq U(x, g^i(x)) + V^i(f(x, g^i(x))) \geq \\ &U(x, h^{i+1}(x)) + V^i(f(x, h^{i+1}(x))) \geq U(x, h^{i+1}(x)) + V^{i+1}(f(x, h^{i+1}(x))) = V^{i+1}(x), \forall x \in \Omega, \end{aligned} \quad (31)$$

The first inequality is due to the assumed  $W^i(x) \geq V^i(x), \forall x$ . The second inequality is due to the fact that  $h^{i+1}(\cdot)$  is the minimizer of the term subject to comparison, and the last inequality is due to the monotonicity of  $\{V^i(x)\}_{i=0}^\infty$ . Hence,  $W^{i+1}(x) \geq V^{i+1}(x), \forall x \in \Omega$ , and the claim is proved by induction.  $\square$

It should be noted that the result given in the former theorem is probably very conservative, as it only shows that the convergence of the PI will not be ‘slower’ than that of the VI.

## VI. CONVERGENCE ANALYSIS OF MULTI-STEP LOOK-AHEAD POLICY ITERATION

Multi-step Look-ahead Policy Iteration (MLPI), [4], is a variation of PI, given by the policy evaluation equation (6) repeated below

$$V^i(x_0) = U(x_0, h^i(x_0)) + V^i(f(x_0, h^i(x_0))), \forall x_0 \in \Omega,$$

and the new *policy update equation* with  $n$ -step look-ahead ( $n \in \mathbb{N}, n > 0$ ) given by

$$h^{i+1}(x_0^h) = \arg \min_{h \in \mathcal{H}} \left( \sum_{k=0}^{n-1} U(x_k^h, h(x_k^h)) + V^i(x_n^h) \right), \forall x_0^h = x_0 \in \Omega. \quad (32)$$

It can be seen that the regular PI is a special case of the MLPI with  $n = 1$ , [4]. It is not surprising to expect the MLPI to converge faster than the regular PI, as in the extreme case that  $n \rightarrow \infty$ , the optimal solution will be calculated in one iteration, using (32), i.e., the iterations converge to the optimal solution after the very first iteration. The rest of this section provides the convergence analysis for MLPI, for  $1 < n < \infty$ .

**Theorem 5.** *The multi-step look-ahead policy iteration given by equations (6) and (32) converges monotonically to the optimal solution in  $\Omega$ .*

*Proof:* The proof is similar to the proof of Theorem 2. Initially it is shown that the sequence of value functions  $\{V^i(x)\}_{i=0}^\infty$  generated using the MLPI is monotonically decreasing. By (32), one has

$$\sum_{k=0}^{n-1} U(x_k^{h^{i+1}}, h^{i+1}(x_k^{h^{i+1}})) + V^i(x_n^{h^{i+1}}) \leq V^i(x_0^{h^{i+1}}), \forall x_0^{h^{i+1}} = x_0 \in \Omega, \quad (33)$$

which is the consequence of  $h^{i+1}(\cdot)$  being the minimizer of the left hand side of the former inequality. Using the line of proof in the proof of Lemma 2, inequality (33) may be repeated in itself for infinite number of times to get

$$V^{i+1}(x) \leq V^i(x), \forall x \in \Omega, \quad (34)$$

which is valid for any selected  $i$ . Hence,  $\{V^i(x)\}_{i=0}^\infty$  under the MLPI is pointwise decreasing. It is also lower bounded by the optimal value function, therefore, converges, [18]. Denoting the limit value function and the limit control policy with  $V^\infty(\cdot)$  and  $h^\infty(\cdot)$ , respectively, they satisfy the MLPI equations

$$V^\infty(x_0) = U(x_0, h^\infty(x_0)) + V^\infty(f(x_0, h^\infty(x_0))), \forall x_0 \in \Omega, \quad (35)$$

and

$$h^\infty(x_0^h) = \arg \min_{h \in \mathcal{H}} \left( \sum_{k=0}^{n-1} U(x_k^h, h(x_k^h)) + V^\infty(x_n^h) \right), \forall x_0^h = x_0 \in \Omega, \quad (36)$$

hence,

$$V^\infty(x_0) = \min_{h \in \mathcal{H}} \left( \sum_{k=0}^{n-1} U(x_k^h, h(x_k^h)) + V^\infty(x_n^h) \right), \forall x_0^h = x_0 \in \Omega. \quad (37)$$

Eq. (37) is the  $n$ -step look-ahead version of the Bellman equation (4) and  $V^*(\cdot)$  satisfies it, by definition. It can be proved that this equation also has a unique solution, which is  $V^*(\cdot)$ , using the line of proof in Lemma 2 and Theorem 1. To this end, assume that

$$V^*(x) < V^\infty(x), \exists x \in \Omega, \quad (38)$$

hence,  $h^*(x) \neq h^\infty(x), \exists x \in \Omega$ . Inequality (38) leads to

$$\sum_{k=0}^{n-1} U(x_k^{h^*}, h^*(x_k^{h^*})) + V^\infty(x_n^{h^*}) < V^\infty(x_0^{h^*}), \exists x_0^{h^*} \in \Omega. \quad (39)$$

Otherwise, one has

$$\sum_{k=0}^{n-1} U(x_k^{h^*}, h^*(x_k^{h^*})) + V^\infty(x_n^{h^*}) \geq V^\infty(x_0^{h^*}), \forall x_0^{h^*} \in \Omega, \quad (40)$$

which repeating it in itself for unlimited number of times, and considering the fact that  $V^\infty(\cdot)$  in the left hand side can be made arbitrarily small after a large enough number of repetitions, leads to

$$V^*(x_0^{h^*}) \geq V^\infty(x_0^{h^*}), \forall x_0^{h^*} \in \Omega. \quad (41)$$

But, (41) contradicts (38), hence, (38) leads to (39). But, (39) contradicts  $h^*(x) \neq h^\infty(x)$ ,  $\exists x \in \Omega$ , per (36), given  $h^*(\cdot) \in \mathcal{H}$ . Therefore, (38) cannot hold and  $V^*(x) = V^\infty(x)$ ,  $\forall x \in \Omega$ , which completes the proof.  $\square$

## VII. CONCLUSIONS

The convergence of the policy iteration scheme to the solution of optimal control problems was analyzed. The speed of convergence of the policy iteration was shown to be not slower than that of the value iteration. Finally, the convergence of the multi-step look-ahead policy iteration to the optimal solution was established.

## REFERENCES

- [1] C. Watkins, *Learning from Delayed Rewards*. PhD Dissertation, Cambridge University, Cambridge, England, 1989.
- [2] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," in *Handbook of Intelligent Control* (D. A. White and D. A. Sofge, eds.), Multiscience Press, 1992.
- [3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.
- [4] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.
- [5] W. B. Powell, *Approximate Dynamic Programming*. Hoboken, NJ, Wiley, 2007.
- [6] S. N. Balakrishnan and V. Biega, "Adaptive-critic based neural networks for aircraft optimal control," *Journal of Guidance, Control and Dynamics*, vol. 19, pp. 893–898, 1996.
- [7] D. Prokhorov and D. Wunsch, "Adaptive critic designs," *IEEE Transactions on Neural Networks*, vol. 8, pp. 997–1007, 1997.
- [8] A. Al-Tamimi, F. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, pp. 943–949, Aug 2008.
- [9] Q. Zhao, H. Xu, and S. Jagannathan, "Optimal control of uncertain quantized linear discrete-time systems," *International Journal of Adaptive Control and Signal Processing*, 2014.
- [10] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 621–634, 2014.
- [11] A. Heydari, "Revisiting approximate dynamic programming and its convergence," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2733–2743, 2014.
- [12] F. Lewis, D. Vrabie, and K. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems*, vol. 32, pp. 76–105, Dec 2012.
- [13] D. P. Bertsekas, "Value and policy iteration in optimal control and adaptive dynamic programming," Report LIDS-P-3174, available online at the authors academic webpage, May 2015.
- [14] D. P. Bertsekas, "Lambda-policy iteration: A review and a new implementation," in *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control* (F. L. Lewis and D. Liu, eds.), pp. 381–406, John Wiley & Sons, 2012.
- [15] H. Khalil, *Nonlinear Systems*. Prentice-Hall, 2002. pp. 111–194.
- [16] R. Howard, *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.
- [17] P. J. Werbos, "Reinforcement learning and approximate dynamic programming (RLADP)-foundations, common misconceptions, and the challenges ahead," in *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control* (F. L. Lewis and D. Liu, eds.), pp. 1–30, John Wiley & Sons, 2012.
- [18] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 3rd ed., 1976. pp. 55, 60, 86, 87, 89, 145, 148.
- [19] B. Lincoln and A. Rantzer, "Relaxing dynamic programming," *IEEE Transactions on Automatic Control*, vol. 51, pp. 1249–1260, Aug 2006.
- [20] A. Heydari, "Stabilizing value iteration with and without approximation errors," 2014. available at arXiv:1412.5675.